UNIVERSITA' DEGLI STUDI DI GENOVA

Data Warehouse: from data to information

Ing. Maurizia Schiozzi

Servizio Statistico, Programmazione e Valutazione

Maurizia.Schiozzi@unige.it

European Commission TEMPUS

UNAM

Agenda

- Case study: University of Genoa
- Data Warehouse and Data Warehousing
- Front-end tools
- Future works

Case study (1)

UNIVERSITY OF GENOA(*)	
Staff (30.11.10)	**2858**
Academic	1433
Administrative	1425
Students (30.11.10)	**34841**
Graduates (01.12.10)	**5442**

(*) More information – **www.unige.it** - “Ateneo **>** Comunicazione **>** Press kit”

Case study (2)

University of Genoa Data warehouse

- 1 application server

  (IBM x336 - 2 CPU 3Ghz, 4GB Ram - 1 volume RAID1 70 GB - Windows 2003 server – Web Server Apache)

- 1 database server

  (IBM P570 - 1 CPU 1.6 Ghz, 2 GB Ram - AIX 5.3 – RDBMS Oracle 9.2.0.7)

- 45 Business Objects Enterprise XI users
- 4 data mart and 25 universes
- 180 tables and 20 Gbyte data

Data mart = subset of data stored in a Data Warehouse, which contains important informations for a particular business area, for a particular division/department, or for particular subjects

Case study (3)

User’s community:

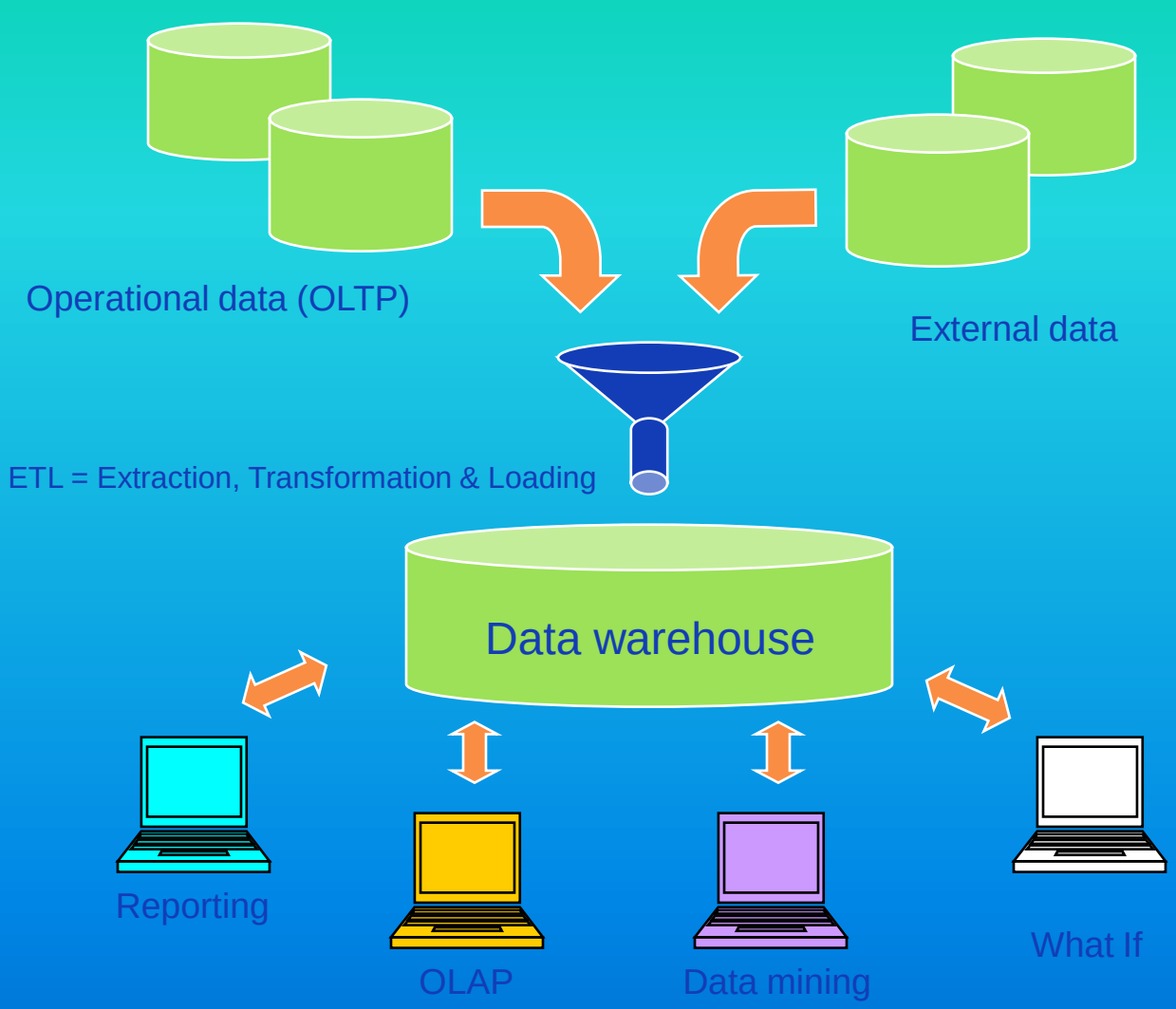
- Governance
- Statistics and evaluation
- Human resources Department
- Financial resources Department
- Student Administration Department
- Faculties

Data warehousing

Data warehouse is a **subject-oriented**, **integrated**, **time-variant** and **non-volatile** collection of data in **support of management's decision making** process *(W.H.Inmon 1990)*.

Data warehousing is the process of constructing and using data warehouse.

Architecture



Sources

Staging

Warehouse

Analysis

OLAP = On Line Analytical Processing

OLTP = On Line Transaction Processing

Data warehouse sources (1)

Data stored in the warehouse come from:

- Operational database of students management software applications
- Operational database of financial management software applications (Campus Accountant)
- Operational database of human resources software applications (Visual Sipert - Byte)
- External data

Data warehouse sources (2)

OLTP – On Line Transaction Processing:

- Database management system are typically used for on-line transaction processing
- OLTP applications normally automate clerical data processing tasks of an organization, like data entry and enqury, transaction handling, etc. (access, read, update)
- Database is current and consistency and recoverabilty are critical. Records are accessed one at a time
- OLTP operations are structured and repetitive
- OLTP operations require detailed and up-to-date data
- OLTP operations are short, atomic and isolated transactions

ETL system

ETL tools populate Data Warehouse.

ETL, or **reconciliation operations,** represent the most difficult and technically complicated operations of the whole data warehousing process.

Reconciliation consists of four phases:

1. Extraction
2. Cleaning
3. Transformation
4. Loading

Staging

Sources data are **extracted**, **cleaned** to eliminate inconsistency and to complete missing parts, **integrated**, to merge heterogeneous sources in a common schema.

ETL (Extraction, Transformation and Loading) systems allow to integrate heterogeneous sources and to **extract**, to **transform**, to **clean**, to **validate**, to **filter** and to **load** data from sources to data warehouse.

Extraction

Only important data are extracted from sources.

Data choice depends on user's particular analysis needs and on data **quality**: because of lack of quality, designers can decide to reject data as they can't provide exact and interesting information.

Cleaning

Cleaning consists of improving data quality by **correction** and **homogeneization** processes to correct errors and inconsistency, for example:

- data duplication
- inconsistency between logically associated values
- missing data
- not expected data use
- impossible or wrong values
- different abbreviations coming from different sources
- typing errors

Transformation

Transformation converts data from operational source format to DW format and it's especially complicated by the presence of multiple distinct and heterogeneous sources.

Some issues need to be corrected at this stage are:

- **presence of free texts** that hide important information
- **use of different formats for the same data** (date)

Loading

Loading data into data warehouse is fixed in time instants, agreed with the user according to the needs of analysis.

Loading is the only operation allowed in a data warehouse, because it is a **read-only** archive.

Data warehouse (1)

A Data Warehouse is a collection of corporate information, derived from operational systems and some external data sources.

Its specific purpose is to ***support business decision****, not business operation.*

Data warehouse (2)

A Data Warehouse is:

- **Subject-oriented** The data in the data warehouse are organized so that all the data elements relating to the same real-world event or object are linked together.
- **Non-volatile** Data in the data warehouse are never over-written or deleted — once committed, the data are static, read-only, and retained for future reporting.
- **Integrated** The data warehouse contains data from most or all of an organization's operational systems and these data are made consistent.
- **Time-variant** In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to operational systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant.

Analysis

Analysis makes data available for efficient and flexible presentation, by **query**, **reporting** and **simulation** (OLAP and data minig tools, What If questions).

In fact, once the data has been cleaned up, integrated, processed and transferred into the data warehouse, designers need to understand **how to take full advantage of information**.

Technically to make a good analysis skills are needed to **explore** aggregated data, and to **optimize** complex queries using **user-friendly** interfaces.

DW architecture (1)

There are two approaches to build a data warehouse :

- ***Top-down***

  To build a core data warehouse first, then use this as the basis to quickly spin off many datamarts: this approach takes longer to build initially since time has to be spent analyzing data requirements in the full warehouse, identifying the data elements that will be used in numerous marts down the road

DW architecture (2)

- ***Bottom-up***

  To build a workgroup specific datamart first: this approach gets data into user's hands quickly, but the work it takes to get the information into the datamart may not be reusable when moving the same data into a warehouse or trying to use similar data in a different datamart.

Datamart building

- Planning
  - Context analysis
  - User requirements analysis
  - Operational data analysis
- Physical environment design
  - Dimensional modeling
  - Data base creation
  - ETL implementation
  - Scheduling
- Logical environment design
  - Design query environment
  - Metadata creation

Planning (1)

Planning is the most delicate phase of the whole project as it is based on specifications supplied by the user.

Once designers understand the user's needs, they have to identify the goals and system boundaries, to estimate the size, cost assessment and value added.

The focus of the project must be as large as possible.

Planning (2)

- **Context analysis**
- Where are the corporate data?
  - Operational database
  - Spreadsheets
  - Other
- How to make data consistent?
- How to transform them into information?
- Logistic and economic considerations

Planning (3)

- **User requirements analysis**
- Identification of expectations
- Identification of system tools and functions
- Consistency between requests and available information
- Application purpose study
- Future needs

Planning (4)

- **Operational data analysis**
  - Check data availability
  - Check quality, consistency and congruence of available data
  - If necessary, introduction of new data in the operational system

Physical environment design (1)

- **Dimensional modeling**

Through the dimensional modeling the designers translate user's requirements into physical structures of data warehouse.

Physical environment design (2)

- **Data base creation**
- Data storage creation
- Structure optimization to reduce the response time of queries

Physical environment design (3)

- **Creation of procedures for extracting data from operational sources and populating the data warehouse**
  Loading is performed by software developed internally by the area of ETL Data Warehouse (PL-SQL language, T-SQL) that perform the following steps :
  - Extraction
  - Transformation (data integration, aggregation, cleaning)
  - Synchronization of data from different sources
  - Time stamping
  - Loading
- **Scheduling**

Logical environment design (1)

Based on the needs of the user, the designers decide how often the data must be photographed by the operational database and loaded into the data warehouse.

According to user's specific requirements, the designers study different types of data presentations, evaluating the need to produce reporting and pre-set navigation paths of interactive data.

Logical environment design (2)

Design query environment (Universe)

The Universe is a semantic layer that contains metadata (data about data) which allows:

- to translate the complexities of the database into business-friendly terms for end-users
- correct SQL generation
- manipulation of data based on the metadata and on the additional business logic in the Universe

Logical environment design(3)

- **Metadata definition**

Metadata describe how and when and by whom a particular set of data was collected, and how the data is formatted. Metadata is essential for understanding information stored in a data warehouse and for translating technical language into business language.

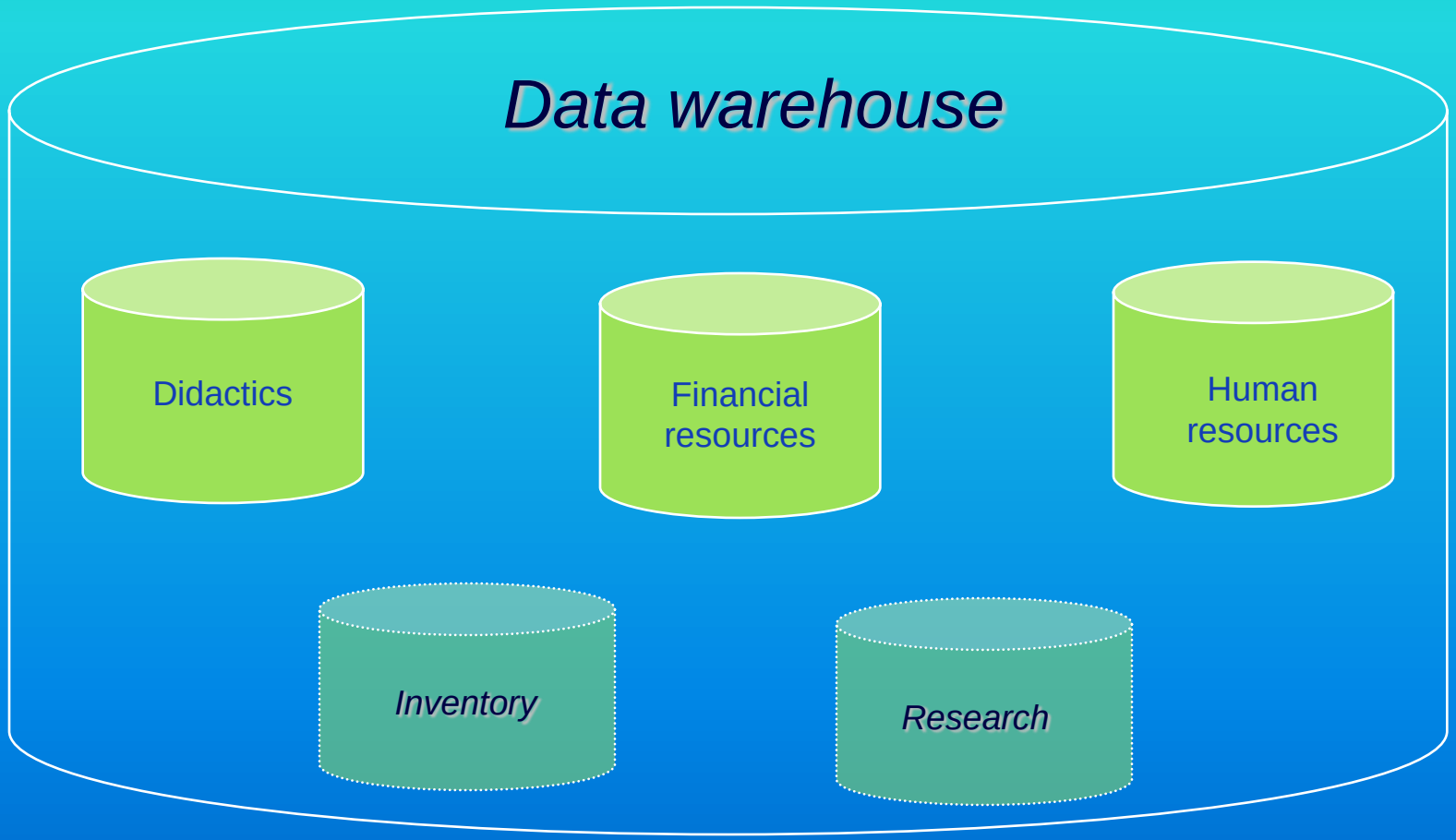
The user’s role

Within a data warehouse project users play an important role and influence all stages of design. In particular, the user **must**:

- provide comprehensive and clear specifications;
- intervene already during the planning phase of the project, to identify goals, assess priorities, and estimate the value added of the construction of a new topic
- intervene on the supply frequency, indicating what is the minimum degree of "freshness of information" it needs
- agree with the designers the common language within the data warehouse, specifying the most appropriate terms for the definition of metadata, the units of measure and the meaningful aggregation
- provide continuous feedback, allowing the designers to evaluate the system in terms of satisfaction and acceptance to the users community

The success of the project depends on customer satisfaction.

University of Genoa Data marts (1)



University of Genoa Data marts (2)

- **Didactics**
  - Undergraduates students
  - Matriculations
  - Graduate students
  - Courses with restricted numbers
  - Study plans
  - Exams
  - Teaching offer
  - Licences to practice
- **Financial resources**
  - Atheneum budget
  - Departments budget
- **Human resources**
  - Academic and administrative staff
  - Salary

Business Objects suite (1)

Business Objects is the suite of products of query, reporting and analysis used by the University of Genoa, which allows to:

- Assure access control and security management
- Define different roles for different user
- Share information
- Take users out of the technical side of data warehouse
- Use and spread a common language
- Access information across the Web
- Create reports
- Analyze information using OLAP operators
- Automatically update and distribute reports

Business Objects suite (2)

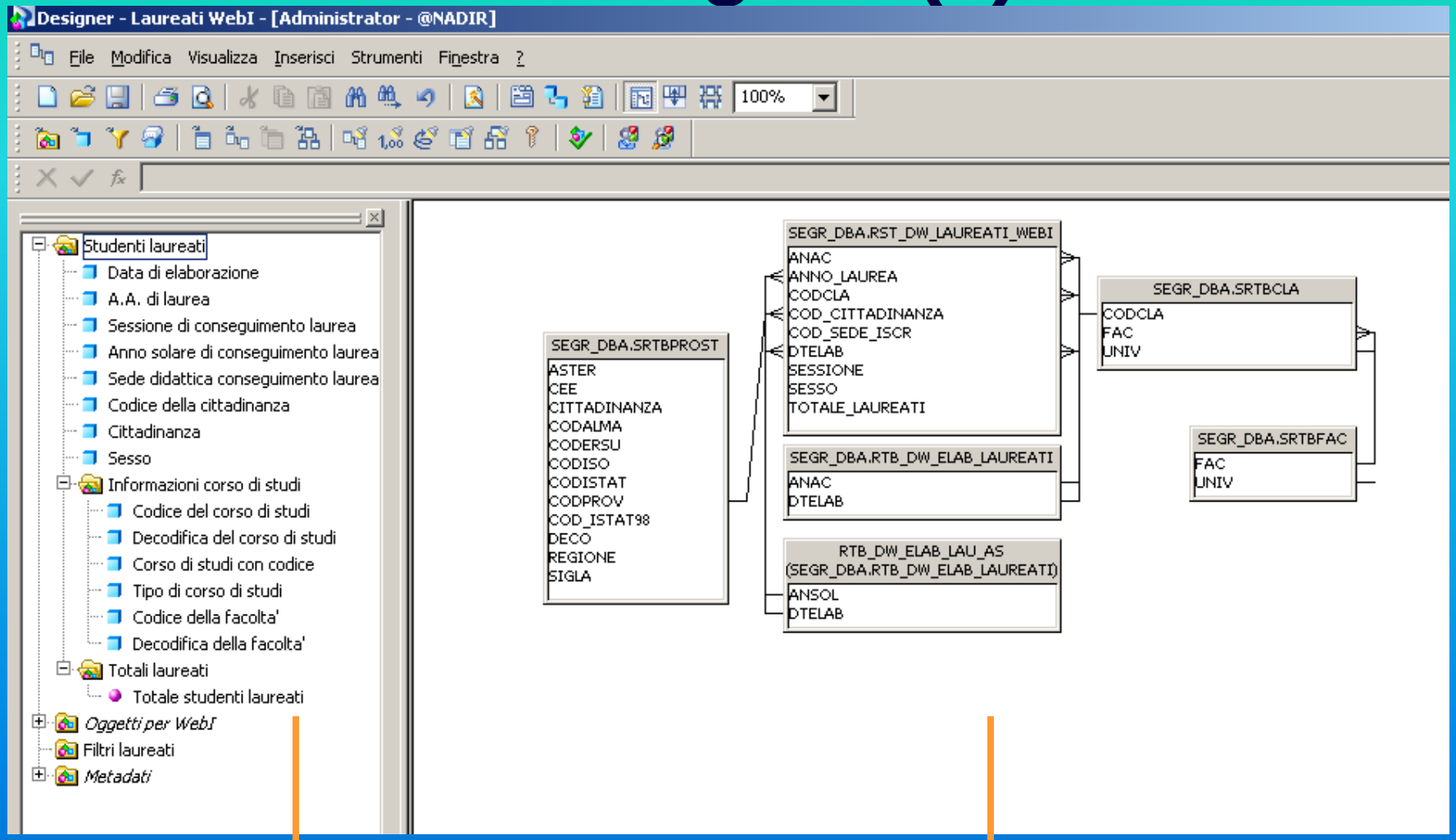
Business Objects Enterprise Edition XI sp 2 includes

- ***Central Mangement Console*** (administrator)
- ***Designer*** (designer)
- ***Desktop Intelligence*** (developer and end user)
- ***Web Intelligence*** (developer and end user)

Designer (1)

- Allows to create Universes, the semantic layers that contains metadata
- Allows to define complex objects
- Permits creation of hyerarchies to explore information at different levels of detail
- Permits metadata definition

Designer (2)



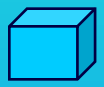
Classes and objects(->query panel)

Database structure

Designer (3)

- **Metadata and universes**

When creating universes, designers define and qualify objects. The qualification of an object reveals how it can be used in analysis in reports.
An object can be qualified as a **dimension** or as a **measure**.



Dimension object

A dimension object is the object being tracked; in other words, it can be considered the focus of the analysis. A dimension can be an object such as Faculty or Course.
Dimension objects retrieve the data that will provide the basis for analysis in a report. Dimension objects typically retrieve character-type data or dates.

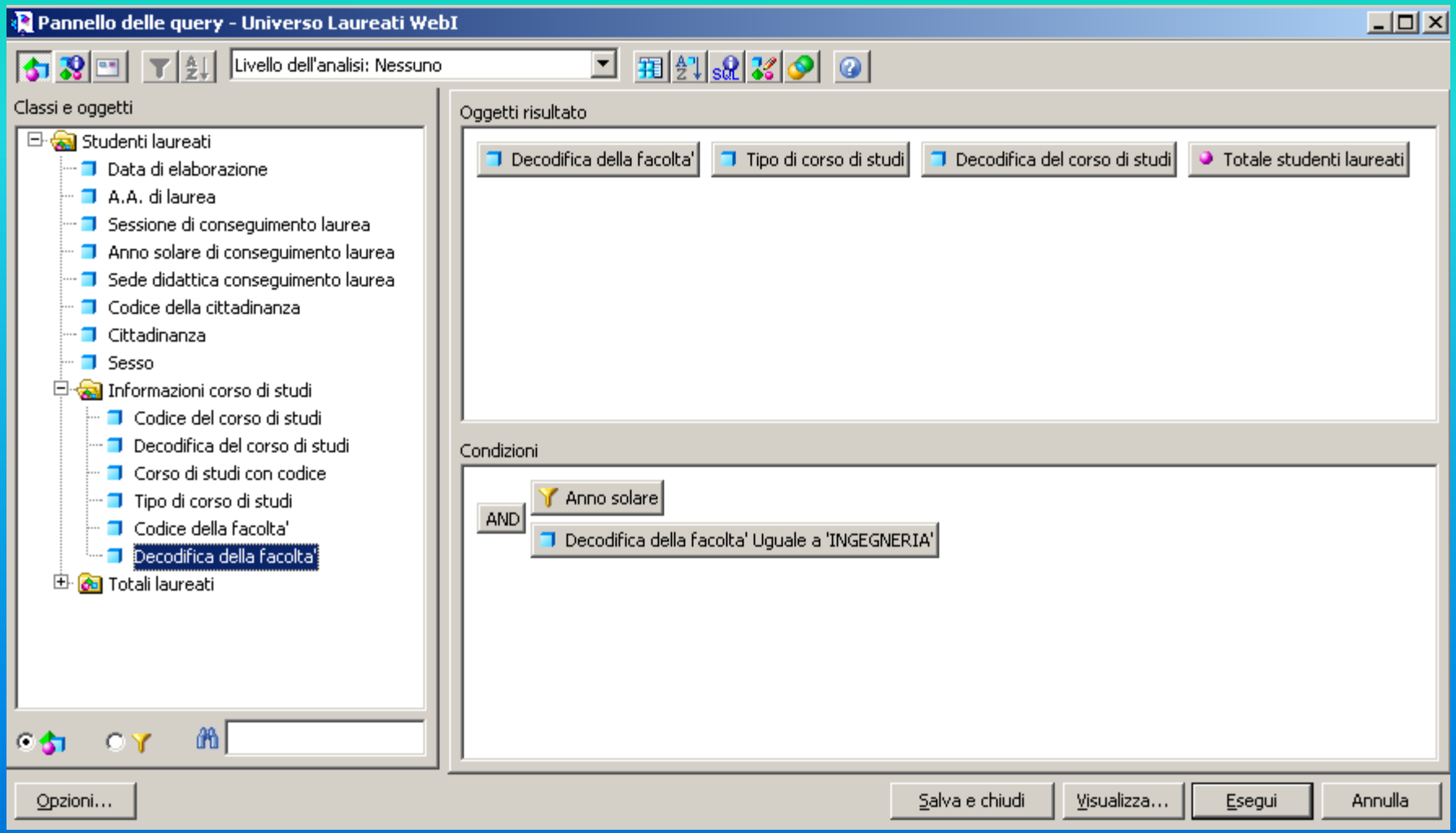
Measure object

A measure object is derived from one of the following aggregate functions: Count, Sum, Minimum, Maximum or average or is a numeric data item on which you can apply, at least locally, one of those functions. This type of object provides statistical information. Examples of measure objects include the following: Number of students, Number of courses….

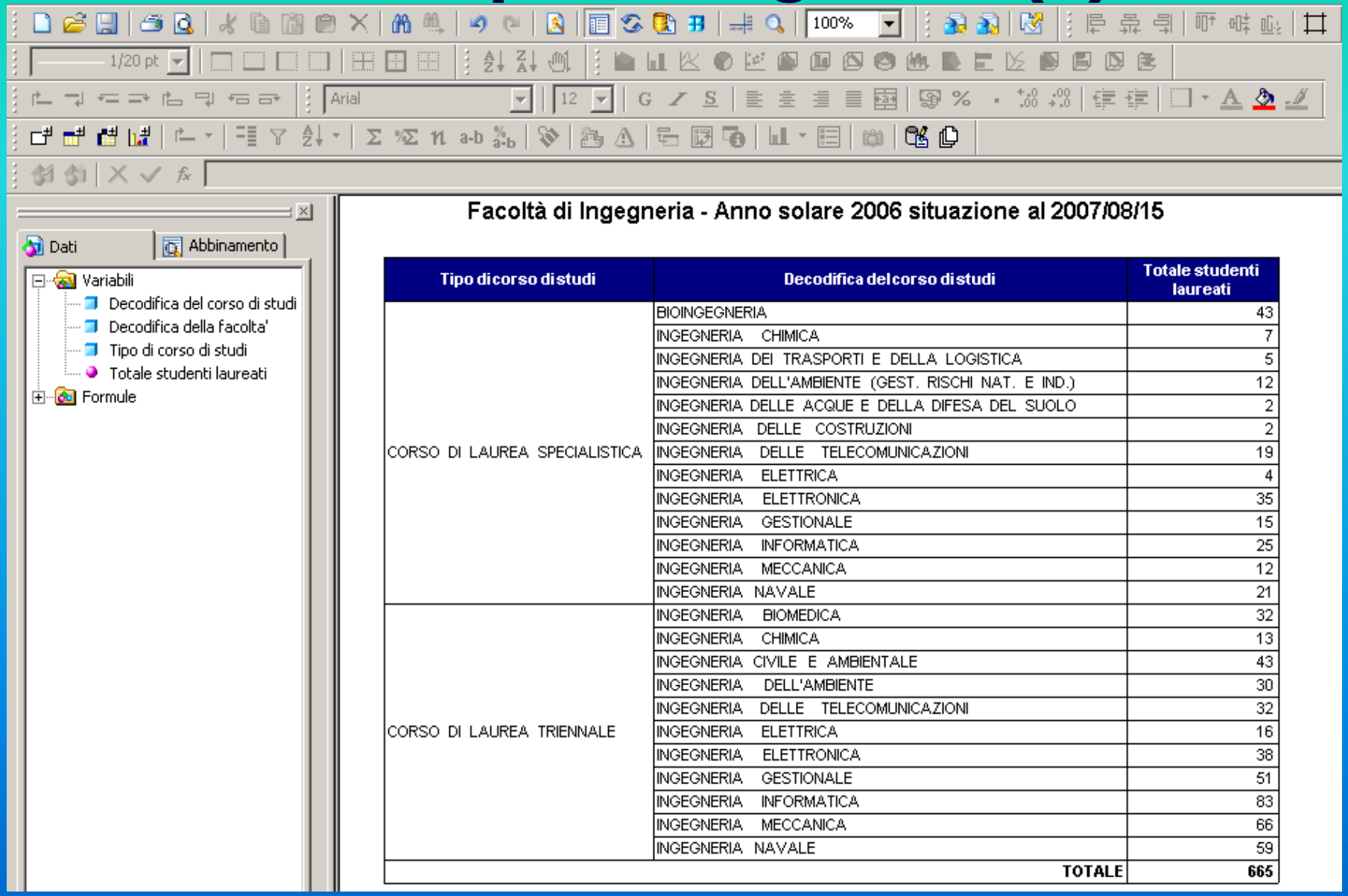
Desktop Intelligence (1)

- Allows to create queries on the universes through the query panel
- Allows to create reports using data coming from
  - Spreadsheet (Excel o ASCII)
  - SQL language
  - Stored procedures or Visual Basic routines
- Permits to organize the result set of a query into dynamic and complex reports, graphs, cross tables: it’s possible to use “pivoting” functions, aggregate functions and to define variables
- Allows to merge queries coming from different universes
- Provides OLAP operators like drill down e slice&dice
- Users can save personal reports, can send them to other users, can publish the reports to the community

Desktop Intelligence (2)



Desktop Intelligence (3)



Facoltà di Ingegneria - Anno solare 2006 situazione al 2007/08/15

Tipo di corso di studi	Decodifica del corso di studi	Totale studenti laureati
CORSO DI LAUREA SPECIALISTICA	BIOINGEGNERIA	43
	INGEGNERIA CHIMICA	7
	INGEGNERIA DEI TRASPORTI E DELLA LOGISTICA	5
	INGEGNERIA DELL'AMBIENTE (GEST. RISCHI NAT. E IND.)	12
	INGEGNERIA DELLE ACQUE E DELLA DIFESA DEL SUOLO	2
	INGEGNERIA DELLE COSTRUZIONI	2
	INGEGNERIA DELLE TELECOMUNICAZIONI	19
	INGEGNERIA ELETTRICA	4
	INGEGNERIA ELETTRONICA	35
	INGEGNERIA GESTIONALE	15
	INGEGNERIA INFORMATICA	25
	INGEGNERIA MECCANICA	12
	INGEGNERIA NAVALE	21
CORSO DI LAUREA TRIENNALE	INGEGNERIA BIOMEDICA	32
	INGEGNERIA CHIMICA	13
	INGEGNERIA CIVILE E AMBIENTALE	43
	INGEGNERIA DELL'AMBIENTE	30
	INGEGNERIA DELLE TELECOMUNICAZIONI	32
	INGEGNERIA ELETTRICA	16
	INGEGNERIA ELETTRONICA	38
	INGEGNERIA GESTIONALE	51
	INGEGNERIA INFORMATICA	83
	INGEGNERIA MECCANICA	66
	INGEGNERIA NAVALE	59
	TOTALE	**665**

Web Intelligence

Web Intelligence is a light version of Business Objects User based on thin client architecture: it allows to build query and create reports using a Web interface, within a common browser, eliminating the need to install and maintain application software on the client and to extablish the connection parameters to the database, ensuring the distribution system outside of the DSS.

- Allows query, reporting and analysis

Future works

- **Design and implementation of new Data Marts (Inventory, Research)**

- **Analysis and implementation of Business Objects Enterprise new functionality**

  - **Dashboard management**

    Dashboard Manager allows to monitor critical business indicators, notify problems that need attention and act on special dashboards to undertake faster decisions.

  - **Performance management**

    Performance Manager allows to create and manage objectives, monitor performances by scorecard, collaborate with other users and undertake the actions suggested to improve business performances.